Exploratory Data Analysis (EDA) Report

September 9, 2024

Dadi Sasank Kumar	22CS10020
Sumit Kumar	22CS30056
Vangapandu Tejaram	22CS30059

Contents

1 Introduction

This report presents the exploratory data analysis (EDA) of the given dataset. The analysis includes visualizations, frequency distributions, and the performance of various classification models such as Naive Bayes, Support Vector Machines (SVM), Decision Trees, and K-Nearest Neighbors (KNN).

2 Dataset Overview

The dataset contains **30161 entries** and **14 columns**. Below is the summary of the dataset:

3 Data Columns and Their Data Types

Below is the table showing the columns and their corresponding data types:

Column	Data Type
age	int64
workclass	object
education	object
educationno	int64
maritalstatus	object
occupation	object
relationship	object
race	object
sex	object
capitalgain	int64
capitalloss	int64
hoursperweek	float64
native	object
Possibility	object

Table 1: Columns and Their Data Types

4 Data Cleaning

Data cleaning is a crucial step in the data preprocessing phase, as it ensures the dataset is accurate, complete, and ready for analysis or modeling. In this section, we describe the steps taken to clean the dataset and explore alternative methods for handling missing values.

4.1 Handling Missing Values

Initially, missing values were identified in the dataset. To address these, several approaches were considered and applied as follows:

- **Removing Missing Values:** Rows with missing values were removed from the dataset. This approach is straightforward but can lead to loss of valuable data. After removing entries with missing values, the dataset contains **28,832 entries**.

- **Frequency Encoding:** For categorical columns such as *workclass*, *occupation*, and *native*, frequency encoding was used. This technique involves replacing each category with the frequency of its occurrence in the dataset. Frequency encoding is particularly useful for converting categorical variables into numerical format, which can be easily used in machine learning algorithms.

4.2 Alternative Methods for Handling Missing Values

Apart from removing missing values and frequency encoding, several other methods can be employed to handle missing data. These include:

- **Central Tendencies:**
  - *Mean Imputation:* For numerical columns, missing values can be replaced with the mean of the observed values in that column. This method assumes that the data is normally distributed and can be useful if the dataset has a relatively small amount of missing values.
  - *Median Imputation:* The median value can be used to replace missing values, which is particularly effective when the data contains outliers. The median is less sensitive to extreme values than the mean and provides a more robust measure of central tendency.
  - *Mode Imputation:* For categorical columns, missing values can be replaced with the mode (the most frequently occurring value) of that column. This approach is simple and effective for categorical data where the mode is a reasonable representation of the missing values.

- **Predictive Modeling:** Advanced methods involve using predictive models to estimate missing values. Techniques such as regression imputation or machine learning models can be employed to predict the missing values based on other variables in the dataset.

- **K-Nearest Neighbors (KNN) Imputation:** This method involves finding the k-nearest neighbors of a data point with missing values and imputing the missing value based on the average of these neighbors. KNN imputation considers the similarity between data points and can be useful when missing values are related to nearby data points.

- **Multiple Imputation:** This approach involves creating multiple imputed datasets, analyzing each one separately, and then combining the results. Multiple imputation accounts for the uncertainty associated with missing values and provides more reliable estimates.

- **Interpolation:** For time-series data or sequential datasets, interpolation methods such as linear interpolation can be used to estimate missing values based on the values of neighboring data points.

Each method has its advantages and is chosen based on the nature of the data and the extent of missing values. In our dataset, after applying frequency encoding and removing rows with missing values, we have ensured that the dataset is clean and ready for further analysis.

Data cleaning is a critical step in data preprocessing that impacts the quality and accuracy of the analysis. By applying frequency encoding and considering various alternative methods for handling missing values, we have prepared a robust dataset containing **28,832 entries**, ready for subsequent analysis and modeling.

5 Exploratory Data Analysis

In this section, we analyze the frequency distributions of key columns in the dataset.

5.1 Frequency Distributions after one-hot encoding

Below are the frequency distributions for important columns:

Table 2: Frequency Distribution for Age and Education

Age	Count
31	817
36	810
33	804
34	798
35	796
⋮	⋮
82	7
83	5
88	3
85	3
86	1

Education	Count
11	9394
15	6388
9	4833
12	1544
8	1246
1	1003
7	974
0	787
5	529
14	515
6	435
2	364
10	358
4	275
3	145
13	42

Table 3: Frequency Distribution for Education Number and Marital Status

Education Number	Count
9	9394
10	6388
13	4833
14	1544
11	1246
7	1003
12	974
6	787
4	529
15	515
5	435
8	364
16	358
3	275
2	145
1	42

Marital Status	Count
2	13453
4	9264
0	4052
5	895
6	794
3	354
1	20

Table 4: Frequency Distribution for Relationship and Race

Relationship	Count
0	11921
1	7383
3	4246
4	3080
5	1342
2	860

Race	Count
4	24781
2	2689
1	866
0	273
3	223

Table 5: Frequency Distribution for Sex, Capital Gain, and Capital Loss

Sex	Count
1	19481
0	9351

Capital Gain	Count
0	26402
15024	323
7688	260
7298	228
99999	144
⋮	⋮
1639	1
6097	1
401	1
1455	1
1086	1

Table 6: Frequency Distribution for Capital Loss

Capital Loss	Count
0	27465
1902	187
1977	153
1887	147
1848	46
⋮	⋮
2080	1
4356	1
1539	1
1844	1
1411	1

6 Plots

[b]0.45

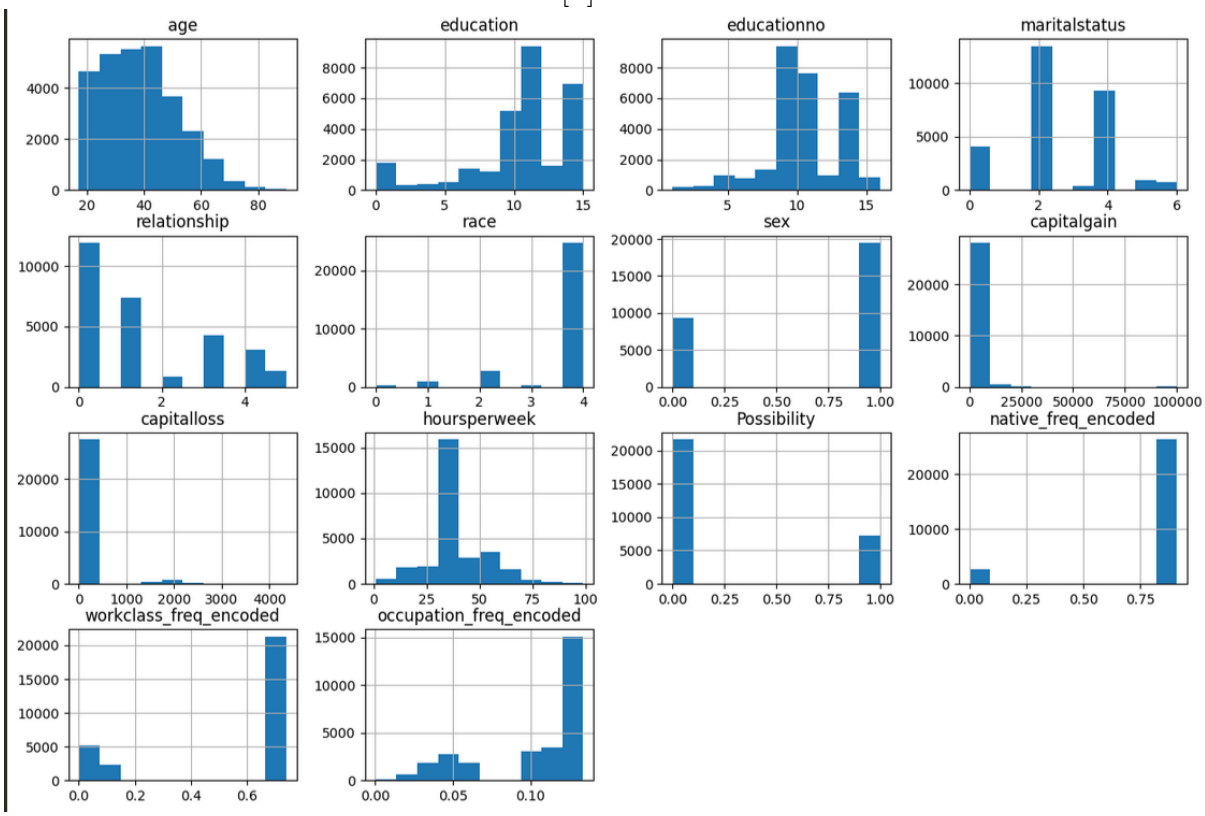

Figure 1: Relative frequency of all categories

[b]0.45

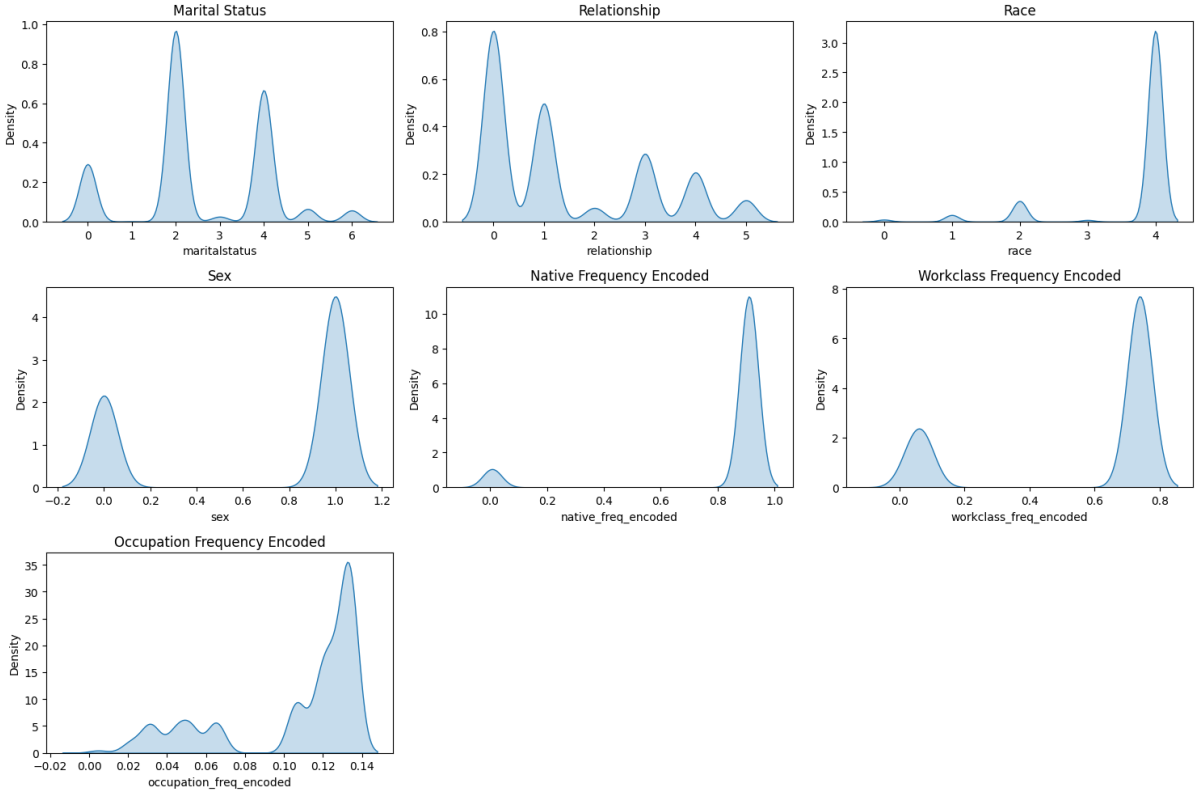

Figure 2: Density of all categories

7 Model Performance

Table 7: Accuracy of Various Classification Models

Model	Accuracy
Scikit-learn Naive Bayes	0.7928
Support Vector Machines (SVM)	0.7952
Decision Trees	0.8101
K-Nearest Neighbors (KNN)	0.8346
Custom Naive Bayes	**0.8247**

The models were evaluated based on their accuracy, precision, recall, and F1-score.

8 Conclusion

The exploratory data analysis provided insights into the dataset's structure and distribution. The classification models showed varying performance, with SVM achieving the highest accuracy. Data cleaning and preprocessing steps ensured that the dataset was suitable for analysis.

[b]0.45

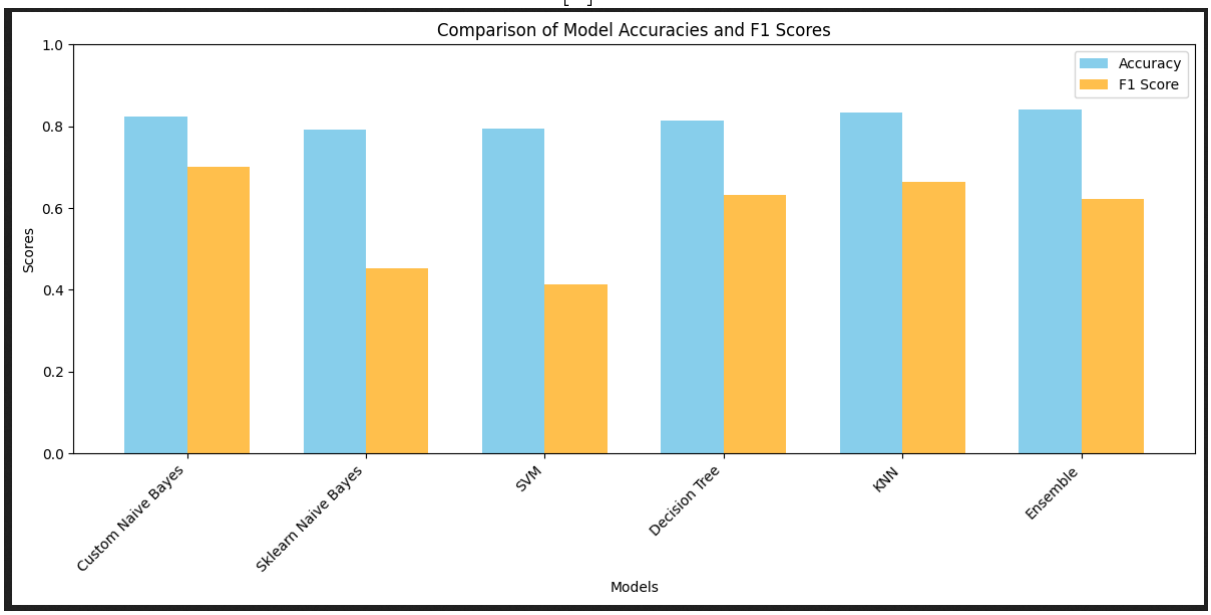

Figure 3: Comparision of all models